

# 第4章 PCI Express シリアル転送におけるパケットの流れ

## ――トランザクション層によるフロー制御の詳細

Steve Kolokowsky，Trevor Davis

**ここでは，PCI Expressアーキテクチャの最上位層にあたるトランザクション層において，データがどのように制御されるかを説明する．PCI Expressでは，あらかじめ受信側のメモリ・バッファ・サイズを確認してから送信を行ったり（クレジット・ベース），必要とする帯域幅に応じて優先順位を付けてパケット送信を行う（仮想チャネル）など，帯域幅やQoS（quality of service）を確保するためのメカニズムが用意されている．（編集部）**

電子機器のメイン・ボードはつねに進化しています．

過去10年以上にわたり，世界中のパソコンや電子機器の開発者は，PCIバス規格を採用し，実装してきました．しかし残念なことに，PCIは次世代の民生機器やパソコン用周辺機器のニーズを満たすことができず，その古い規格ゆえの欠点を露呈し始めました．その結果，「PCIからPCI Expressへ移行すれば問題が解決する」と考える設計者が増え続けています．

## 1 ビデオ・カードから始まった

PCIでは帯域幅が足りず問題となった最初の周辺機器は，ハイエンドのビデオ・カードです．

この問題を解消するため，業界はAGP（Accelerated Graphics Port）を専用ビデオ・バスとして採用しました．AGPの提唱者たちは，米国Intel社のICH5チップセットにあるメモリ・コントローラ（MCH）にバスを直結することで，最終的にAGPの帯域幅を2 Gバイト/sまで押し上げました．

AGPで始まった帯域幅拡大の動きとともに，ほかの周辺機器でも十分な帯域幅を得るためにオーディオやUSBインターフェースなど，あらゆる機能が直接メイン・ボードのチップセットに接続されるようになりました．その結果，パソコンのアーキテクチャとして，ギガビットEthernetのような高速な帯域幅に対応できる新しい拡張バスが必要となりました．

### ● グラフィックス・カードは8Gバイト/sのx16を採用

帯域幅の要求にこたえるため，PCIの高速・広帯域バージョンや，一部のサーバ設計に採用されたPCI-Xなど，いくつかの標準バス規格が現れては消えていきました．こうした中，Intel社のメイン・ボード用チップセットの仕様の変更を受けて，PCI-SIG（Special Interest Group）がPCI Express（当時の名称は「3GIO」）を採用することが確実となりました．

PCI Expressは拡張性を備えたバスです．現在の仕様（Rev. 1. x）では，2.5GHzのバス・クロックで250Mバイト/s～8Gバイト/sの帯域幅に対応します．クロック速度やバス幅を上げることで，さらなる高速化も可能です．PCI Expressのもっとも重要な技術的メリットの一つに，PCI用ソフトウェアとの下位互換性があります．これによって，既存のBIOSやOSといっしょに利用できます（**表1**）．

また，広帯域幅を必要とするデバイスの場合，2/4/8/12/16/32とレーン数を増やすことができます．16レーン（x16）は，次世代のグラフィックス・バス・インターフェースとして導入され始めています．また，PCI-SIG内で策定されているPCI Expressのケーブル仕様が完成すれば，PCI Expressを使用して広帯域幅の周辺機器をパソコンの外部に接続できるようになります．

パソコンにおけるPCIと同じように，PCI Expressはサウス・ブリッジの周辺機器の接続バスとして実装されます．

また，PCI Expressはグラフィックスの用途ではノース・ブリッジから拡張することも可能です．**図1**に，パソコンにおける低コストのPCI Expressの実装例を示します．現在のところ，パソコン周辺機器の接続にはPCIバスが多く採用されていますが，帯域幅の拡大に伴い，PCI Expressへ急速に移行するものと予測されます．

**表1　PCI Expressの帯域幅とバスの互換性**

| バス規格 | 帯域幅（Mバイト/s） |
|---|---|
| PCI（バス幅：32ビット，バス動作周波数：33MHz） | 133 |
| PCI（バス幅：64ビット，バス動作周波数：33MHz） | 266 |
| PCI Express（x1，双方向） | 500 |
| PCI Express（x2，双方向） | 1,000 |
| PCI Express（x4，双方向） | 2,000 |
| PCI Express（x8，双方向） | 4,000 |
| PCI Express（x12，双方向） | 6,000 |
| PCI Express（x16，双方向） | 8,000 |

## 2 実際のデータの流れ

次に，PCI Expressのデータの流れを理解するため，典型的なトランザクションについて解説します．

### ● データの読み出し/書き込みなどの処理を行う最上位層

PCI Expressのバス・トラフィックは，ほかのバス・プロトコルと同じようにパケットを利用してデバイス間の情報通信を行っています．パケットはトランザクション層とデータリンク層で構築され，送信側のデバイスから受信側のデバイスへ情報が伝達されます（**図2**）．送信されたパケットは相手側の層に入り，そこで変更されてパケットを処理するのに必要な情報が追加されます．

物理層には，処理に必要なインターフェースの電気回路ブロック，およびデバイス間のインターフェースの初期化や保守に必要な論理機能が含まれています．さらに，物理層はデータリンク層から受信した情報をシリアル形式に変換し，リンクの反対側に接続されたデバイスと互換性のある周波数やバス幅でPCI Expressリンク全体に送信する役割も果たします．

データリンク層は，トランザクション層と物理層の中継点の役割を果たします．そのおもな役割は，エラーの検出や修正を含むデータの整合性，リンクの管理などです．

PCI Expressアーキテクチャの最上位にあるのがトランザクション層です．トランザクション層は，トランザクションの通信に使用するTLP（Transaction Layer Packet）の組み立てと分解，TLPのクレジット・ベースのフロー制御という二つの役割を果たします．

### ● クレジット・ベースの制御フローで帯域幅を確保

ここで，PCI Expressのトラフィック（パケットの流れ）について，例を挙げて説明します．

例えば，Webブラウザなどのアプリケーションで，Ethernetカードから受信したデータを読み出す処理（トラ

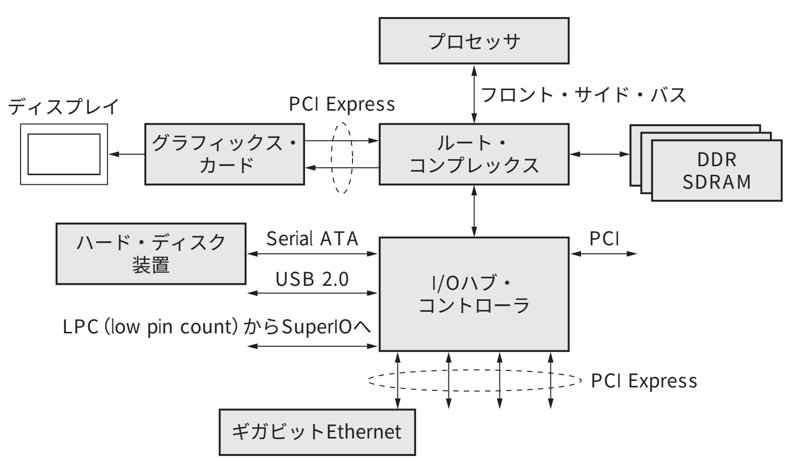



**図1　PCI Express搭載の低コスト・アーキテクチャ**
ルート・コンプレックスは，PCI ExpressバスとCPUを接続するデバイス．通常，ルート・コンプレックスは複数ポートをサポートしているが，必要なポートは一つのみ．広帯域幅のグラフィックス・カードに対応するため，x16の単一ポートを実装したノース・ブリッジもある．

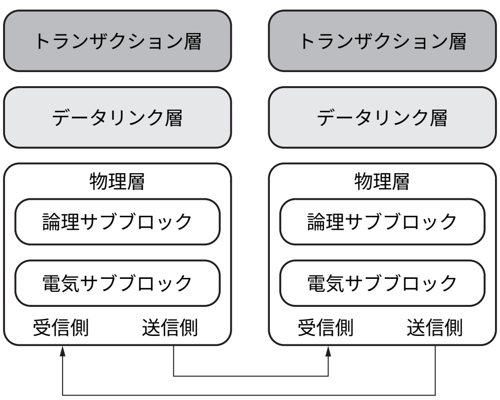



**図2　PCI Expressのトランザクション層**
アーキテクチャの最上位に位置するのがトランザクション層．PCI Expressで使用するトランザクションの種類は，「ポステッド」，「ノンポステッド」，「コンプリーション」の三つ．ポステッド・トランザクションは，応答を必要としないメモリ書き込みのような処理．一方，ノンポステッド・トランザクションにはターゲットからのコンプリーション・トランザクションが必要であり，例えばメモリ読み取りがこれに相当する．

ンザクション）が必要であるとします．このトランザクションを実行するには，ブラウザでメモリ・バッファを読み出すための要求（リクエスト）を生成します．この要求はソフトウェアのいくつかの層に伝達され，最終的にHAL（ハードウェア分離層）への経路が見つかります．HALはWindowsの一部で，論理アドレスを物理アドレスに変換するものです．要求がHALを経由すると，CPUが物理メモリ・アドレスの読み出しを実行します．

読み出しトランザクションがCPUを通過すると，フロント・サイド・バスを経由してノース・ブリッジに渡されます．ノース・ブリッジではトランザクションのアドレスに注目し，アドレスがサウス・ブリッジを指しているかどうかを確認してからトランザクションを転送します．サウス・ブリッジには，PCI Expressに対する初回の読み出し要求を移行するためのPCI Expressスイッチがあります．

この時点で，読み出し要求がルート・コンプレックスのキューに入れられます（**図3**）．読み出し要求がキューの最前列に到達すると，ルート・コンプレックスによってこの要求をTLP内に配置します．これで，リンク経由でパケットを送信する準備は完了です．ただし，まず，受信側のデバイスにパケットを保持するのに十分なメモリ容量があるかどうかを，ルート・コンプレックスに確認する必要があります．

PCI Expressは，受信側の使用可能なバッファ・サイズを調べてから送信する「クレジット・ベースのフロー制御」という方式を採用しています．エンドポイントまたはスイッチが起動すると，リンク接続先へDLLP（Data Link Layer Packet）を送信します．DLLPは，トランザクションの種類ごとにどれくらいのバッファ空間を使用できるかを知らせるものです．リトライ・ベースのフロー制御では，使用可能な空間を用いてシステムが受信バッファを検索するため，大量の帯域幅を急速に消費してしまいます．

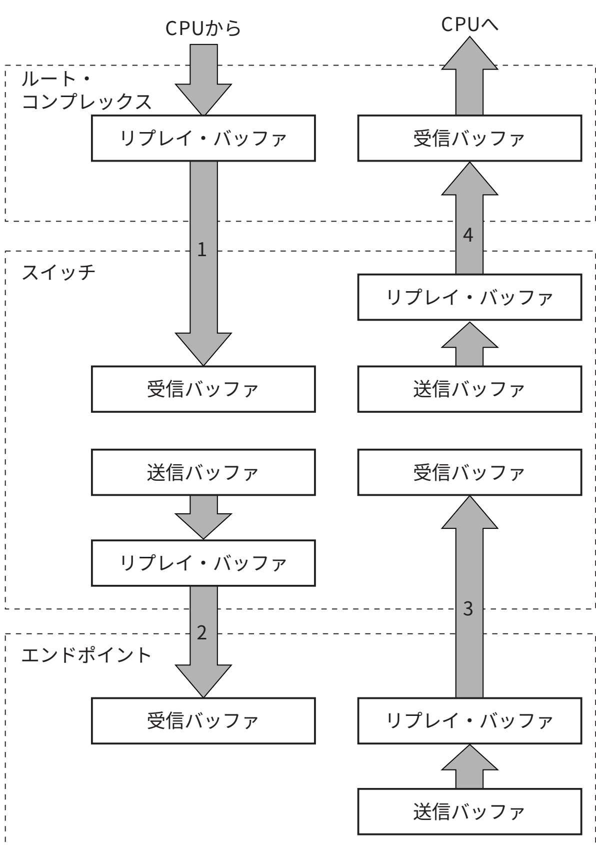



**図3　PCI Expressのトランザクション・フロー**

**読み出し要求がキューの最前列に到達すると，ルート・コンプレックスでこの要求をTLP内に配置する．ここで，メモリ読み出し/書き込み，I/Oの読み出し/書き込みなどのパケットはすべてTLPである．図の中のスイッチは，USBハブのような役割を果たし，複数の機器を単一のリンクに接続するためのデバイス．**

### ● 優先順位を付けてフロー制御を行う

PCI Expressの新機能に，仮想チャネル（VC：virtual channel）という概念があります．仮想チャネルにより，優先順位の高いトラフィックを低いものよりも先に各リンクに送信できます．優先順位の設定はリンクの両側で行われます．送信側に送信待ちのパケットが複数ある場合，優先順位の高い仮想チャネル・パケットから送信するように選択できます．受信側は，仮想チャネルごとに個別のフロー制御カウンタを持つ専用バッファを備えています．

仮想チャネルは，トラフィック・クラス（必要に応じた帯域の割り当て）と緊密に連携しています．トラフィック・クラスはトラフィックの優先順位を示し，これを仮想チャネルが利用します．例えば，トラフィック・クラスを速達郵便，普通郵便，小包といった複数の種類の郵便物に，仮想チャネルをその配送車と考えてください．配送車が3台以上あれば，速達郵便をいちばん速い車両で，普通郵便を次に速い車両で，小包をもっとも遅い車両で配達します．しかし，配送車が1台しかない場合は，すべての郵便物を同じ優先順位でいっしょに配達することになります．

**図3**の場合，ルート・コンプレックスに適切なトランザクションの種類と仮想チャネルの十分なクレジットがあれば，リンク経由で読み出し要求を送信し（**図3**の矢印1），パケットのコピーがリプレイ・バッファに格納されます．次に，スイッチが読み出し要求を受信し，アドレス・フィ

ールドを確認します．さらに，アドレスを参照して適切なダウンストリーム・ポートを決定します．保留中のメッセージがほかにある場合，スイッチから優先順位の高いトラフィックが先に送信され，その後，デバイスに読み出し要求が転送されます(**図3**の矢印2)．

読み出し要求を受信すると，デバイスはその要求を受信バッファに格納します．この時点でも，優先順位の高い仮想チャネルから先に渡される可能性があります．受信バッファが空の場合，要求が直ちに処理されます．次に，読み出し完了パケットがデバイスの転送バッファのキューに入れられ，スイッチを経由してルート・コンプレックスに戻されます．サウス・ブリッジで読み出し完了パケットを受信すると，ノース・ブリッジを経由してCPUに戻され，トランザクションが完了します．

### ● 実際のバス・トラフィックを解析する

ここまでで，実際のデータの流れを検証するために必要なPCI Expressのバス・トラフィックについては理解できたと思います．

**図4**に，米国LeCroy社のPCI Expressプロトコル解析ソフトウェア「CATC Trace」[注]の表示をもとに作成した図を示します．本ソフトウェアには，状況を明確にするためのいくつかのフィールドが含まれています．パケット番号の隣にある矢印はバスの方向を示しており，“R←”はアップストリームを，“R→”はダウンストリームを意味します．アップストリームおよびダウンストリームのデータ経路は，それぞれ同時にアクティブにできるパラレルのパスですが，**図4**では両者を合わせて一つのトレースとみなしています．

その隣のフィールドは，パケットの種類(TLPまたはDLLP)を示しています．残りのフィールドにはパケットのペイロードが含まれていて，現在のパケットと次のパケットのデータ経路における遅延を表しています．

ここで，それぞれのパケットを詳しく見てみましょう．

- パケット0000001 ── ここでのトランザクションは，アドレスDFCFCE68に対して1 DW(double word；1DW＝32ビットとする)のメモリ書き込み要求を行うホストによって開始される．これは，ポステッド・トランザクション(TLPでは応答を必要としないトランザクション．ただし，DLLPでは確認応答が必要)である
- パケット0000002 ── ホストの次のパケットは，アドレスDFCFCF28のメモリ読み出しである．メモリの読み出しはノンポステッド・トランザクションなので，メモリ・データを含むTLPレベルでの完了応答が必要となる．パケット0000001の段のトレース画面に「Idle 0.000 ns」と示されているが，これはパケット0000002がメモリ書き込みの直後にあることを示している．PCI

4

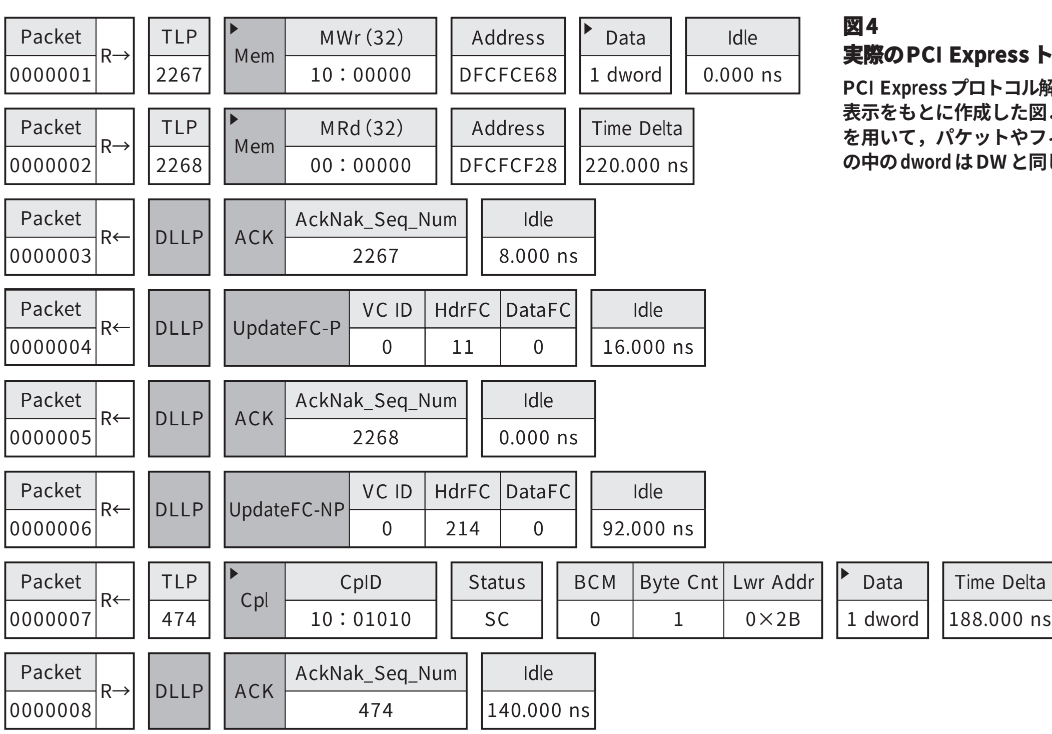



**図4**
**実際のPCI Expressトラフィック**
PCI Expressプロトコル解析ソフトウェア「CATC Trace」の表示をもとに作成した図．こうしたプロトコル解析ツールを用いて，パケットやフィールドの確認を行う．なお，図の中のdwordはDWと同じ意味である．

注：LeCroy社は，2004年10月に本ソフトウェアの開発元であるバス・アナライザ・メーカの米国CATC(Computer Access Technology Corp.)を買収した．

**表2 フロー制御トラフィックの詳細**

**バス・トラフィックのシーケンスと通知された状態を示す．**

| 実行中動作 | トラフィック・ステータス | | | 備　考 |
|---|---|---|---|---|
| — | ←Init-FC-10 | | | フロー制御の初期設定．バッファ・サイズとして10を開放 |
| バッファ・ステータス1 | Start | End | Available | |
| | 0 | 10 | 10 | この時点のTLPではバッファを使用しない |
| 次のバス・トラフィック | →TLP-size 3 | | | TLPとしてバッファ・サイズ3を利用 |
| バッファ・ステータス2 | Start | End | Available | |
| | 3 | 10 | 7 | 利用可能なバッファ・サイズは7（＝10－3） |
| 次のバス・トラフィック | →TLP-size 2 | | | TLPとしてさらにバッファ（2）を利用 |
| バッファ・ステータス3 | Start | End | Available | |
| | 5 | 10 | 5 | 利用可能なバッファ・サイズは5 |
| 次のバス・トラフィック | ←Update-FC-13 | | | バッファ・サイズの更新（10→13） |
| バッファ・ステータス4 | Start | End | Available | |
| | 5 | 13 | 8 | バッファ・サイズの更新により利用可能な空間は8（＝13－5） |
| 次のバス・トラフィック | →TLP-size 4 | | | TLPとしてさらにバッファ（4）を利用 |
| バッファ・ステータス5 | Start | End | Available | |
| | 9 | 13 | 4 | この時点で利用可能なバッファ・サイズは4 |

Expressでは，クレジット・ベースのフロー制御によって，ホストは書き込み後に受信側に残っているバッファ空間の確認を待たなくてもパケットを送信できる．また，番号付きのTLPパケットにより，ホストは前のTLPのACK（受信応答）を待たなくても別のTLPを送信できる

- パケット0000003～0000006 ── 次の四つのパケットは，エンドポイントからルート・コンプレックスへのDLLPパケット．パケット0000003と0000005はACKで，対応するパケット（およびそれよりも前にあるすべてのTLPパケット）が正しく受信されたことを示す．パケット0000004と0000006は，使用可能なバッファ空間ではなく，現在のバッファの位置を示すことにより，クレジット・ベースのフロー制御を更新する．この二つは，送信中でもバッファ空間に関する正確な状況を送信側で把握するために使用する
- パケット0000007 ── このパケットは，パケット0000002の読み出し要求の完了（コンプリーション）を含む
- パケット0000008 ── 最後のパケットには，パケット0000007のACKが含まれている．この時点で，パケット0000007で使用するバッファ空間を開放する「Update FC」メッセージ以外に，保留中のDLLPは存在しない

### ● フロー制御が成功へのカギ

仮想チャネルとフロー制御のメカニズムはお互いに関係があります．仮想チャネルはデータの優先順位を設定するのに対して，フロー制御は受信バッファのオーバフローを防ぎ，仮想チャネルの優先順位設定の規則に準拠するために使用します．

最初になんらかのリンクが設定されると，各エンドポイントがトラフィックの種類と仮想チャネルごとに「Init-FC」メッセージを送信します（**表2**）．仮想チャネルが2個あるシステムの場合，フロー制御により6個のバッファが存在することになります．そのうち三つは仮想チャネル0のポステッド・バッファ，ノンポステッド・バッファ，コンプリーション・バッファで，残り三つは仮想チャネル1の3種類のトラフィックです．この「Init-FC」メッセージは，送信側に使用可能なバッファ空間がどのくらいあるかを通知します．それ以降は，「Update-FC」メッセージが使用可能なクレジットの数ではなく，現在のバッファの位置を送信側に通知します．

**表2**のStart列はリセット時に‘0’に初期化され，その後，送信されるフロー制御用のクレジット数が保持されます．End列は，受信側バッファの最上位を示します．Available列は，「End列－Start列」で算出される受信側で使用可能なバッファ空間を示します．

PCI Expressは，いくつかの複雑なメカニズムを採用しています．これらのメカニズムは，高速なデータ転送を維持するクレジット・ベースのフロー制御，QoSを高め，かつバッファ・サイズを縮小するために帯域幅の優先順位を設定する仮想チャネルなど，データをできる限り高速かつ確実に送信するためのものです．こうした複雑なメカニズムを理解することで，システム内に適切な通信経路を設定できる敏腕システム設計者になれるでしょう．

**Steve Kolokowsky，Trevor Davis**
**米国Cypress Semiconductor社**